Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization

Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization

Ömer Deniz Akyildiz[⋆,†] and Sotirios Sabanis[⋆,‡]

[⋆]The Alan Turing Institute, London, UK.
[†]University of Warwick, Coventry, UK.
[‡]University of Edinburgh, UK.
omer.akyildiz@warwick.ac.uk, s.sabanis@ed.ac.uk

February 14, 2020

Abstract

We provide a nonasymptotic analysis of the convergence of the stochastic gradient Hamiltonian Monte Carlo (SGHMC) to a target measure in Wasserstein-2 distance without assuming log-concavity. By making the dimension dependence explicit, we provide a uniform convergence rate of order $\mathcal{O}(\eta^{1/4})$, where η is the step-size. Our results shed light onto the performance of the SGHMC methods compared to their overdamped counterparts, e.g., stochastic gradient Langevin dynamics (SGLD). Furthermore, our results also imply that the SGHMC, when viewed as a nonconvex optimizer, converges to a global minimum with the best known rates.

1 Introduction

We are interested in nonasymptotic estimates for the sampling problem from the probability measures of the form

$$\pi_\beta(\mathrm{d}\theta) \propto \exp(-\beta U(\theta))\mathrm{d}\theta. \tag{1}$$

when only the noisy estimate of ∇U is available. This problem arises in many cases in machine learning, most notably in large-scale (mini-batch) Bayesian inference (Welling and Teh, 2011, Ahn et al., 2012) and nonconvex stochastic optimization (Raginsky et al., 2017). For the setting of Bayesian inference, one is interested in sampling from a posterior probability measure where U corresponds to the sum of the log-likelihood and the log-prior. For the nonconvex optimization, $U(\cdot)$ is the nonconvex cost function to be minimized. For large values of β, a sample from the target measure (1) is an approximate minimizer of the potential U (Raginsky et al., 2017). Consequently, nonasymptotic error bounds for the schemes, which are designed to sample from (1), can be used to obtain guarantees for Bayesian inference or nonconvex optimization. Sampling from a measure of the form (1) is also central in statistical physics (Binder et al., 1993), most notably in molecular dynamics Haile (1992).

An efficient method for obtaining a sample from (1) is simulating the overdamped Langevin stochastic differential equation (SDE) which is given by

$$\mathrm{d}L_t = -h(L_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t, \tag{2}$$

with a random initial condition $L_0 := \theta_0$ where $h := \nabla U$ and $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. The Langevin SDE (2) admits π_β as the unique invariant measure, therefore simulating this

This work was supported by The Alan Turing Institute for Data Science and AI under EPSRC grant EP/N510129/1. Ö. D. A. is funded by the Lloyds Register Foundation programme on Data Centric Engineering through the London Air Quality project.

process will lead to samples from π_β and can be used as a Markov chain Monte Carlo (MCMC) algorithm (Roberts et al., 1996, Roberts and Stramer, 2002). Moreover, the fact that the limiting probability measure π_β concentrates around the global minimum of U for sufficiently large values of β makes the diffusion (2) also an attractive candidate as a global optimizer (see, e.g., Hwang (1980)). However, since the continuous-time process (2) can not be simulated, its first-order Euler discretization with the step-size $\eta > 0$ is used in practice, termed *the Unadjusted Langevin Algorithm* (ULA) (Roberts et al., 1996). The ULA scheme has become popular in recent years due to its advantages in high-dimensional settings and ease of implementation. Nonasymptotic properties of the ULA were recently established under strong convexity and smoothness assumptions in Dalalyan (2017), Durmus et al. (2017, 2019) while some extensions about relaxing smoothness assumptions or inaccurate gradients were also considered in Dalalyan and Karagulyan (2019), Brosse et al. (2019). The similar attractive properties hold for the ULA when the potential U is nonconvex (Gelfand and Mitter, 1991, Raginsky et al., 2017, Xu et al., 2018, Erdogdu et al., 2018, Sabanis and Zhang, 2019).

While the ULA performs well when the computation of the gradient $h(\cdot)$ is straightforward, this is not the case in most interesting applications. Usually, a stochastic, unbiased estimate of $h(\cdot)$ is available, either because the cost function is defined as an expectation or as a finite sum. Using stochastic instead of deterministic gradients in the ULA leads to another scheme called stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011). The SGLD has been particularly popular in the fields of (i) *large-scale Bayesian inference* since it allows one to construct Markov chains Monte Carlo (MCMC) algorithms using only subsets of the dataset (Welling and Teh, 2011), (ii) *nonconvex optimization* since it enables one to estimate global minima using only stochastic (often cheap-to-compute) gradients (Raginsky et al., 2017). As a result, attempts for theoretical understanding of the SGLD have been recently made in several works, both for the strongly convex potentials (i.e. log-concave targets), see, e.g., (Barkhagen et al., 2018, Brosse et al., 2018) and nonconvex potentials, see, e.g. Raginsky et al. (2017), Majka et al. (2018), Zhang et al. (2019). Our particular interest is in nonasymptotic bounds for nonconvex case, as it is relevant to our work. We note that the seminal paper Raginsky et al. (2017) obtains a nonasymptotic bound between the law of the SGLD and the target measure in Wasserstein-2 distance with a rate $\eta^{5/4}n$ where η is the step-size and n is the number of iterations. While this work is first of its kind, the error rate grows with the number of iterations. In a related contribution, Xu et al. (2018) have obtained improved rates, albeit still growing with the number of iterations n. In more recent work, Chau et al. (2019b) have obtained a uniform rate of order $\eta^{1/2}$ in Wasserstein-1 distance. Majka et al. (2018) achieved error rates of $\eta^{1/2}$ and $\eta^{1/4}$ for Wasserstein-1 and Wasserstein-2 distances, respectively, under the assumption of convexity at infinity. Finally, Zhang et al. (2019) achieved the same rates under only local conditions which can be verified for a class of practical problems.

An alternative to the methods based on the *overdamped* Langevin SDE (2) is the class of algorithms which are based on the *underdamped* Langevin SDE. To be precise, the underdamped Langevin SDE is given as

$$\mathrm{d}V_t = -\gamma V_t \mathrm{d}t - h(\theta_t)\mathrm{d}t + \sqrt{\frac{2\gamma}{\beta}}\mathrm{d}B_t, \tag{3}$$

$$\mathrm{d}\theta_t = V_t \mathrm{d}t, \tag{4}$$

where $(\theta_t, V_t)_{t\geq 0}$ are called position and momentum process, respectively, and $h := \nabla U$. Similar to eq. (2), this diffusion can be used as both an MCMC sampler and nonconvex optimizer, since under appropriate conditions, the Markov process $(\theta_t, V_t)_{t\geq 0}$ has a unique invariant measure given by

$$\overline{\pi}_\beta(\mathrm{d}\theta, \mathrm{d}v) \propto \exp\left(-\beta\left(\frac{1}{2}\|v\|^2 + U(\theta)\right)\right)\mathrm{d}\theta\mathrm{d}v. \tag{5}$$

Consequently, the marginal distribution of (5) in θ is precisely the target measure defined in (1).

This means that sampling from (5) in the extended space and then keeping the samples in the θ-space would define a valid sampler for the sampling problem of (1).

Due to its attractive properties, methods based on the underdamped Langevin SDE have attracted significant attention. In particular, the first order discretization of (3)–(4), which is termed *underdamped Langevin MCMC* (i.e. the *underdamped* counterpart of the ULA), has been a focus of attention, see, e.g., Duncan et al. (2017), Dalalyan and Riou-Durand (2018), Cheng et al. (2018b). Particularly, the underdamped Langevin MCMC has displayed improved convergence rates in the setting where U is convex, see, e.g., Dalalyan and Riou-Durand (2018), Cheng et al. (2018b). Similar results have been extended to the nonconvex case. In particular, Cheng et al. (2018a) have shown that the underdamped Langevin MCMC converges in Wasserstein-2 with a better dimension and step-size dependence under the assumptions smoothness and convexity outside a ball. It has been also shown that the underdamped Langevin MCMC can be seen as an accelerated optimization method in the space of measures in Kullback-Leibler divergence (Ma et al., 2019) which partially explains its improved convergence properties.

Similar to the case in the ULA, oftentimes $\nabla U(\cdot)$ is expensive or impossible to compute exactly, but rather an unbiased estimate of it can be obtained efficiently. When one replaces the gradient in the underdamped Langevin MCMC with a stochastic gradient, the resulting method is dubbed as Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) and given as

$$V_{n+1}^{\eta} = V_n^{\eta} - \eta[\gamma V_n^{\eta} + H(\theta_n^{\eta}, X_{n+1})] + \sqrt{\frac{2\gamma\eta}{\beta}}\xi_{n+1}, \tag{6}$$

$$\theta_{n+1}^{\eta} = \theta_n^{\eta} + \eta V_n^{\eta}, \tag{7}$$

where $\eta > 0$ is a step-size, $V_0^{\eta} = v_0$, $\theta_0^{\eta} = \theta_0$, and $\mathbb{E}[H(\theta, X_0)] = h(\theta)$ for every $\theta \in \mathbb{R}^d$.

In this paper, we analyze recursions (6)–(7). We achieve convergence bounds and improve the existing ones proved in Gao et al. (2018) and Chau and Rasonyi (2019) (see Section 2.1 for a direct comparison).

Notation. For an integer $d \geq 1$, the Borel sigma-algebra of $\mathbb{R}^d$ is denoted by $\mathcal{B}(\mathbb{R}^d)$. We denote the dot product with $\langle \cdot, \cdot \rangle$ while $|\cdot|$ denotes the associated norm. The set of probability measures defined on a measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is denoted as $\mathcal{P}(\mathbb{R}^d)$. For an $\mathbb{R}^d$-valued random variable, $\mathcal{L}(X)$ and $\mathbb{E}[X]$ are used to denote its law and its expectation respectively. Note that we also write $\mathbb{E}[X]$ as $\mathbb{E}X$ when there is no risk of confusion. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures Γ on $\mathcal{B}(\mathbb{R}^{2d})$ so that its marginals are μ, ν. Finally, we define the Wasserstein distance of order $p \geq 1$ as

$$W_p(\mu, \nu) := \inf_{\Gamma \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta - \theta'|^p \Gamma(\mathrm{d}\theta, \mathrm{d}\theta') \right)^{1/p}, \tag{8}$$

for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.

2 Main results and overview

In this section, our main theoretical results are introduced and a detailed comparison is given with the most recent findings in the literature.

Let $(X_n)_{n\in\mathbb{N}}$ be an $\mathbb{R}^m$-valued stochastic process adapted to $(\mathcal{G}_n)_{n\in\mathbb{N}}$ where $\mathcal{G}_n := \sigma(X_k, k \leq n, k \in \mathbb{N})$ for $n \in \mathbb{N}$. It is assumed henceforth that $\theta_0, v_0, \mathcal{G}_\infty$, and $(\xi_n)_{n\in\mathbb{N}}$ are independent. The main assumptions follow.

Assumption 2.1. *The cost function U takes nonnegative values, i.e., $U(\theta) \geq 0$.*

Despite the fact that we restrict our attention to nonnegative potentials, we note that this case covers a large number of applications in nonconvex optimization and sampling. The following assumption states that the stochastic gradients are assumed to be unbiased.

Assumption 2.2. *The process* $(X_n)_{n\in\mathbb{N}}$ *is i.i.d. with* $|X_0| \in L^{4(\rho+1)}$ *and* $|\theta_0|, |v_0| \in L^4$. *It satisfies*

$$\mathbb{E}[H(\theta, X_0)] = h(\theta).$$

Next, the requirements on the stochastic gradients $H(\theta, \cdot)$ are given, in particular with respect to their local smoothness properties.

Assumption 2.3. *There exist positive constants* L_1, L_2 *and* ρ *such that, for all* $x, x' \in \mathbb{R}^m$ *and* $\theta, \theta' \in \mathbb{R}^d$,

$$|H(\theta, x) - H(\theta', x)| \le L_1(1 + |x|)^\rho |\theta - \theta'|,$$
$$|H(\theta, x) - H(\theta, x')| \le L_2(1 + |x| + |x'|)^\rho (1 + |\theta|)|x - x'|.$$

It is important to note that Assumption 2.3 is a significant relaxation in comparison with the corresponding assumptions provided in the literature, see, e.g., Raginsky et al. (2017), Gao et al. (2018), Chau and Rasonyi (2019). To the best of the authors' knowledge, all relevant works in this area have focused on uniform Lipschitz assumptions with the exception of Zhang et al. (2019), which provides a nonasymptotic analysis of the SGLD under similar assumptions to ours. Next, we present an important remark following from Assumption 2.3.

Remark 2.1. *Assumption 2.3 implies, for all* $\theta, \theta' \in \mathbb{R}^d$,

$$|h(\theta) - h(\theta')| \le L_1 \mathbb{E}[(1 + |X_0|)^\rho]|\theta - \theta'|, \tag{9}$$

which consequently implies

$$|h(\theta)| \le L_1 \mathbb{E}[(1 + |X_0|)^\rho]|\theta| + h_0, \tag{10}$$

where $h_0 := |h(0)|$. *Also, Assumption 2.3 implies*

$$|H(\theta, x)| \le L_1(1 + |x|)^\rho |\theta| + L_2(1 + |x|)^{\rho+1} + H_0.$$

where $H_0 := |H(0, 0)|$.

We denote

$$C_\rho := \mathbb{E}\left[(1 + |X_0|)^{4(\rho+1)}\right].$$

Note that $C_\rho < \infty$ by Assumption 2.2.

We present a dissipativity assumption next.

Assumption 2.4. *There exist* $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$, $b : \mathbb{R}^m \to \mathbb{R}$ *such that for any* $x, y \in \mathbb{R}^d$,

$$\langle y, A(x)y \rangle \ge 0$$

and for all $\theta \in \mathbb{R}^d$ *and* $x \in \mathbb{R}^m$,

$$\langle H(\theta, x), \theta \rangle \ge \langle \theta, A(x)\theta \rangle - b(x).$$

The smallest eigenvalue of $\mathbb{E}[A(X_0)]$ *is a positive real number* $a > 0$ *and* $\mathbb{E}[b(X_0)] = b > 0$.

Note that Assumption 2.4 is a *local* condition. This assumption implies the usual dissipativity property on the corresponding deterministic (full) gradient.

Remark 2.2. *By Assumption 2.4, we obtain*

$$\langle h(\theta), \theta \rangle \ge a|\theta|^2 - b,$$

for $\theta \in \mathbb{R}^d$ *and* $a, b > 0$.

Below, we state our main result about the convergence of the law $\mathcal{L}(\theta_k^\eta, V_k^\eta)$, which is generated by the SGHMC recursions (6)–(7), to the extended target measure $\overline{\pi}_\beta$ in Wasserstein-2 (W_2) distance. We first define

$$\eta_{\max} = \min\left\{\frac{K_3}{K_2}, \frac{\gamma\lambda}{K_1}, \frac{1}{\gamma\lambda}\right\}$$

where

$$K_1 := \max\left\{\frac{2(2L_1 C_\rho + \gamma^2 - \gamma^2\lambda + \gamma)}{(1-2\lambda)}, \frac{8(\widetilde{L}_1 + \gamma L_1 C_\rho^2)}{(1-2\lambda)\gamma^2}\right\}$$

and

$$K_2 = \frac{\widetilde{C_1} + \gamma h_0^2}{2} \quad \text{and} \quad K_3 = (A_c + d)\gamma\beta^{-1}$$

Then, the following result is obtained.

Theorem 2.1. *Let Assumptions 2.1, 2.2, 2.3 and 2.4 hold. Then, there exist constants $C_0^\star, C_1^\star, C_2^\star, C_3^\star > 0$ such that, for every $0 < \eta \leq \eta_{\max}$,*

$$W_2(\mathcal{L}(\theta_n^\eta, V_n^\eta), \overline{\pi}_\beta) \leq C_1^\star \eta^{1/2} d^{1/2} + C_2^\star \eta^{1/4} d^{3/4} + C_3^\star e^{-C_4^\star \eta n}. \tag{11}$$

We note that although $C_1^\star$ is a dimension free constant, $C_2^\star$ and $C_3^\star$ may exhibit exponential dependence on the dimension d as it is an immediate consequence of the contraction result of the underdamped Langevin SDE in Eberle et al. (2019). Moreover, the obtained rate demonstrates that the error scales like $\mathcal{O}(\eta^{1/4})$ and is uniformly bounded over n which can be made arbitrarily small by choosing $\eta > 0$ small enough. This result is thus a significant improvement over the findings in Gao et al. (2018), where error bounds are presented that grow with the number of iterations, and in Chau and Rasonyi (2019), where the corresponding error bounds contain an additional term that is independent of η and relates to the variance of the unbiased estimator.

Remark 2.3. *We note that our proof techniques can be adapted easily when $H(\theta, x) = h(\theta)$ hence Theorem 2.1 provides a convergence rate for the analysis of the underdamped Langevin MCMC under our relaxed assumptions which itself is a novel contribution.*

Let $(\theta_k^\eta)_{k\in\mathbb{N}}$ be generated by the SGHMC algorithm. Convergence of the $\mathcal{L}(\theta_k^\eta)$ to π_β in W_2 also implies that one can prove a global convergence result (Raginsky et al., 2017). More precisely, assume that we aim at solving the problem

$$\theta_\star \in \arg\min_{\theta\in\mathbb{R}^d} U(\theta) \tag{12}$$

which is a nonconvex optimization problem. We denote $U_\star := \inf_{\theta\in\mathbb{R}^d} U(\theta)$. Then we can bound the error $\mathbb{E}[U(\theta_k^\eta)] - U_\star$ which would give us a guarantee on the nonconvex optimization problem. We state it as a formal result as follows.

Theorem 2.2. *Under the assumptions of Theorem 2.1, we obtain*

$$\mathbb{E}[U(\theta_n^\eta)] - U_\star \leq \overline{C}_1^\star \eta^{1/2} d^{1/2} + \overline{C}_2^\star \eta^{1/4} d^{3/4} + \overline{C}_3^\star e^{-C_4^\star \eta n} + \frac{d}{2\beta}\log\left(\frac{e\overline{L}_1}{a}\left(\frac{b\beta}{d}+1\right)\right),$$

where $\overline{C}_1^\star, \overline{C}_2^\star, \overline{C}_3^\star, C_4^\star, \overline{L}_1$ are finite constants which are explicitly given in the proof.

This result bounds the error in terms of the function value for convergence to the global minima.

2.1 Related work and contributions

Our work is most related to two available analyses of the SGHMC, namely Gao et al. (2018) and Chau and Rasonyi (2019). We contrast the convergence rates provided in Theorem 2.1 and 2.2 to these two works.

The scheme (6)–(7) is analyzed in Gao et al. (2018). In particular, Gao et al. (2018) provided a convergence rate of the SGHMC (6)–(7) to the underdamped Langevin SDE (3)–(4) which is of order $\mathcal{O}(\delta^{1/4} + \eta^{1/4})\sqrt{n\eta}\sqrt{\log(\eta n)}$. This rate grows with n, hence worsens over the number of iterations. Moreover, it is achieved under a uniform assumption on the stochastic gradient, i.e., $H(\theta, x)$ is assumed to be Lipschitz in θ uniformly in x (as opposed to our Assumption 2.3). Moreover, the mean-squared error of the gradient is assumed to be bounded whereas we do not place such an assumption in our work. Similar analyses appeared in the literature, e.g., for variance-reduced SGHMC (Zou et al., 2019) which also has growing rates with the number of iterations.

Another related work was provided by Chau and Rasonyi (2019) who also analyzed the SGHMC recursions essentially under the same assumptions as in Gao et al. (2018). However, Chau and Rasonyi (2019) improved the convergence rate of the SGHMC recursions to the underdamped Langevin SDE significantly, i.e., provided a convergence rate of order $\mathcal{O}(\delta^{1/4} + \eta^{1/4})$ where $\delta > 0$ is a constant. While this rate significantly improves the rate of Gao et al. (2018), it cannot be made to vanish by choosing $\eta > 0$ small enough, as $\delta > 0$ is (a priori assumed to be) independent of η.

In contrast, we prove that the SGHMC recursions track the underdamped Langevin SDE with a rate of order $\mathcal{O}(\eta^{1/4})$ which can be made arbitrarily small as with small $\eta > 0$. Moreover, our assumptions are significantly relaxed compared to Gao et al. (2018) and Chau and Rasonyi (2019). In particular, we relax the assumptions on stochastic gradients significantly by allowing growth in both variables (θ, x) which makes our theory hold for practical settings such as variational inference (Zhang et al., 2019).

3 Preliminary results

In this section, preliminary results which are essential for proving the main results are provided. A central idea behind the proof of Theorem 2.1 is the introduction of continuous-time auxiliary processes whose marginals at chosen discrete times coincide with the marginals of the (joint) law $\mathcal{L}(\theta_k^\eta, V_k^\eta)$. Hence, these auxiliary stochastic processes can be used to analyze the theoretical properties of the recursions (6)–(7).

3.1 Introduction of the auxiliary processes

We first define the scaled process $(\zeta_t^\eta, Z_t^\eta) := (\theta_{\eta t}, V_{\eta t})$ where $(\theta_t, V_t)_{t\in\mathbb{R}_+}$ are defined as in (3)–(4). We next define

$$\mathrm{d}Z_t^\eta = -\eta(\gamma Z_t^\eta + h(\zeta_t^\eta))\mathrm{d}t + \sqrt{2\gamma\eta\beta^{-1}}\mathrm{d}B_t^\eta, \tag{13}$$
$$\mathrm{d}\zeta_t^\eta = \eta Z_t^\eta \mathrm{d}t, \tag{14}$$

where $\eta > 0$ and $B_t^\eta = \frac{1}{\sqrt{\eta}} B_{\eta t}$, where $(B_s)_{s\in\mathbb{R}_+}$ is a Brownian motion with natural filtration $\mathcal{F}_t$. We denote the natural filtration of $(B_t^\eta)_{t\in\mathbb{R}_+}$ as $\mathcal{F}_t^\eta$. We note that $\mathcal{F}_t^\eta$ is independent of $\mathcal{G}_\infty \vee \sigma(\theta_0, v_0)$. Next, we define the continuous-time interpolation of the SGHMC

$$\mathrm{d}\overline{V}_t^\eta = -\eta(\gamma\overline{V}_{\lfloor t\rfloor}^\eta + H(\overline{\theta}_{\lfloor t\rfloor}^\eta, X_{\lceil t\rceil}))\mathrm{d}t + \sqrt{2\gamma\eta\beta^{-1}}\mathrm{d}B_t^\eta, \tag{15}$$
$$\mathrm{d}\overline{\theta}_t^\eta = \eta\overline{V}_{\lfloor t\rfloor}^\eta \mathrm{d}t. \tag{16}$$

It is easy to verify that the processes (15)–(16) mimic the recursions (6)–(7) at discrete times $n \in \mathbb{N}$, i.e., $\mathcal{L}(\theta_n^\eta, V_n^\eta) = \mathcal{L}(\overline{\theta}_n^\eta, \overline{V}_n^\eta)$.

Finally, we define the underdamped Langevin process $(\widehat{\zeta}_t^{s,u,v,\eta}, \widehat{Z}_t^{s,u,v,\eta})$ for $s \le t$

$$\mathrm{d}\widehat{Z}_t^{s,u,v,\eta} = -\eta(\gamma \widehat{Z}_t^{s,u,v,\eta} + h(\widehat{\zeta}_t^{s,u,v,\eta}))\mathrm{d}t + \sqrt{2\gamma\eta\beta^{-1}}\mathrm{d}B_t^\eta, \tag{17}$$

$$\mathrm{d}\widehat{\zeta}_t^{s,u,v,\eta} = \eta \widehat{Z}_t^{s,u,v,\eta}\mathrm{d}t, \tag{18}$$

with initial conditions $\widehat{\theta}_s^{s,u,v,\eta} = u$ and $\widehat{V}_s^{s,u,v,\eta} = v$. This process is a regular underdamped Langevin SDE which is started at points (u, v).

Definition 3.1. *Fix $n \in \mathbb{N}$ and define*

$$\overline{\zeta}_t^{\eta,n} = \widehat{\zeta}_t^{nT,\overline{\theta}_{nT}^\eta,\overline{V}_{nT}^\eta,\eta},$$

$$\overline{Z}_t^{\eta,n} = \widehat{Z}_t^{nT,\overline{\theta}_{nT}^\eta,\overline{V}_{nT}^\eta,\eta},$$

where $T := \lfloor 1/\eta \rfloor$.

The process $(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})_{t \ge nT}$ is the underdamped Langevin process started at time nT with initial data $(\overline{\theta}_{nT}^\eta, \overline{V}_{nT}^\eta)$.

3.2 Moment estimates and contraction rates

To achieve the convergence results, we first define a Lyapunov function, borrowed from Eberle et al. (2019) as

$$\mathcal{V}(\theta, v) = \beta U(\theta) + \frac{\beta}{4}\gamma^2 \left(\|\theta + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|\theta\|^2 \right). \tag{19}$$

where $\lambda \in (0, 1/4]$. This Lyapunov function plays an important role in obtaining uniform moment estimates for some of the aforementioned processes. Next, it is shown that a key assumption appearing in Eberle et al. (2019) holds.

Lemma 3.1. *There exist constants $A_c \in (0, \infty)$ and $\lambda \in (0, 1/4]$ such that*

$$\langle \theta, h(\theta) \rangle \ge 2\lambda \left(U(\theta) + \gamma^2 |\theta|^2/4 \right) - 2A_c/\beta \tag{20}$$

for all $x \in \mathbb{R}^d$.

Further, uniform in time, second moment estimates for θ_t^η and V_t^η are obtained, in view of Remarks 2.1 and 2.2.

Lemma 3.2. *(Lemma 12(i) in Gao et al. (2018).)Let Assumptions 2.1–2.4 hold. Then*

$$\sup_{t \ge 0} \mathbb{E}|\theta_t|^2 \le C_\theta^c := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{d + A_c}{\lambda}}{\frac{1}{8}(1 - 2\lambda)\beta\gamma^2}, \tag{21}$$

$$\sup_{t \ge 0} \mathbb{E}|V_t|^2 \le C_v^c := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{d + A_c}{\lambda}}{\frac{1}{4}(1 - 2\lambda)\beta} \tag{22}$$

Moreover, an analogous result holds true also for the discrete-time processes $(\theta_k^\eta)_{k \ge 0}$ and $(V_k^\eta)_{k \ge 0}$.

Lemma 3.3. *Let Assumptions 2.1–2.4 hold. Then, for $0 < \eta \le \eta_{\max}$,*

$$\sup_{k \ge 0} \mathbb{E}|\theta_k^\eta|^2 \le C_\theta := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{2(A_c + d)}{\lambda}}{\frac{1}{8}(1 - 2\lambda)\beta\gamma^2},$$

$$\sup_{k \ge 0} \mathbb{E}|V_k^\eta|^2 \le C_v := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta, v)\mu_0(\mathrm{d}\theta, \mathrm{d}v) + \frac{2(A_c + d)}{\lambda}}{\frac{1}{4}(1 - 2\lambda)\beta}.$$

As a result, uniform moment estimates are obtained for $\overline{\zeta}_t^{\eta,n}$ when $t \ge nT$.

Lemma 3.4. *Under the assumptions of Lemmas 3.2 and 3.3, we obtain*

$$\sup_{n\in\mathbb{N}}\sup_{t\in(nT,(n+1)T)}\mathbb{E}|\overline{\zeta}_t^{\eta,n}|^2\leq C_\zeta, \tag{23}$$

where

$$C_\zeta := \frac{\int_{\mathbb{R}^{2d}}\mathcal{V}(\theta,v)\mu_0(\mathrm{d}\theta,\mathrm{d}v)+\frac{4(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}.$$

Finally, an exponential convergence rate for the underdamped Langevin diffusion is presented in accordance to the findings in Eberle et al. (2019). To this end, a functional for probability measures μ,ν on $\mathbb{R}^{2d}$ is introduced below

$$\mathcal{W}_\rho(\mu,\nu)=\inf_{\Gamma\in\mathcal{C}(\mu,\nu)}\int\rho((x,v),(x',v'))\Gamma(d(x,v),d(x',v')), \tag{24}$$

where ρ is defined in eq. (2.10) in Eberle et al. (2019). Thus, in view of Remarks 2.1 and 2.2, one recovers the following result.

Theorem 3.1. *(Eberle et al. (2019) Theorem 2.3 and Corollary 2.6) Let Assumptions 2.1–2.4 hold and the laws of the underdamped Langevin SDEs (θ_t,V_t) and (θ'_t,V'_t) started at $(\theta_0,V_0)\sim\mu$ and $(\theta'_0,V'_0)\sim\nu$ respectively. Then, there exist constants $\dot{c},\dot{C}\in(0,\infty)$ such that*

$$W_2(\mathcal{L}(\theta_t,V_t),\mathcal{L}(\theta'_t,V'_t))\leq\sqrt{\dot{C}}e^{-\dot{c}t/2}\sqrt{\mathcal{W}_\rho(\mu,\nu)}, \tag{25}$$

where

$$\dot{C}=2e^{2+\Lambda}\frac{(1+\gamma)^2}{\min(1,\alpha)^2}\max\left(1,4(1+2\alpha+2\alpha^2)\frac{(d+A_c)\gamma^{-1}\dot{c}^{-1}}{\min(1,R_1)\beta}\right)$$

and

$$\dot{c}=\frac{\gamma}{384}\min\left(\lambda\overline{L}_1\beta^{-1}\gamma^{-2},\Lambda^{1/2}e^{-\Lambda}\overline{L}_1\beta^{-1}\gamma^{-2},\Lambda^{1/2}e^{-\Lambda}\right)$$

where

$$\Lambda=\overline{L}_1R_1^2/8=\frac{12}{5}(1+2\alpha+2\alpha^2)(d+A_c)\overline{L}_1\beta^{-1}\gamma^{-2}\lambda^{-1}(1-2\lambda)^{-1},$$

and $\alpha\in(0,\infty)$.

We note that, due to the contraction result Theorem 3.1 of Eberle et al. (2019), the dimension dependence of $\dot{C}$ may be exponential.

4 Proof of Theorem 2.1

In order to prove Theorem 2.1, we note first that W_2 is a metric on $\mathcal{P}(\mathbb{R}^{2d})$. The main strategy for this proof is to bound $W_2(\mathcal{L}(\theta_n^\eta,V_n^\eta),\overline{\pi}_\beta)$ by using appropriate estimates on the continuous time interpolation of $(\theta_n^\eta,V_n^\eta)_{n\in\mathbb{N}}$. In particular, we obtain the desired bound by decomposing first in such a way

$$\begin{aligned}W_2(\mathcal{L}(\overline{\theta}_t^\eta,\overline{V}_t^\eta),\overline{\pi}_\beta)\leq{}&W_2(\mathcal{L}(\overline{\theta}_t,\overline{V}_t^\eta),\mathcal{L}(\overline{\zeta}_t^{\eta,n},\overline{Z}_t^{\eta,n}))+W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n},\overline{Z}_t^{\eta,n}),\mathcal{L}(\zeta_t^\eta,Z_t^\eta))\\&+W_2(\mathcal{L}(\zeta_t^\eta,Z_t^\eta),\overline{\pi}_\beta),\end{aligned} \tag{26}$$

and then by obtaining suitable (decaying in η) bounds for each of the three terms appearing on the right hand side of (26). This leads to the proof of our main result, namely, Theorem 2.1.

We bound the first term of (26) as follows.

Theorem 4.1. *Let Assumptions 2.1–2.4 hold and* $0 < \eta \leq \eta_{\max}$. *Then,*

$$W_2(\mathcal{L}(\overline{\theta}_t^\eta, \overline{V}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})) \leq C_1^\star d^{1/2} \eta^{1/2} \tag{27}$$

where $C_1^\star < \infty$ *and independent of* d.

Next, we prove the following result for bounding the second term of (26).

Theorem 4.2. *Let Assumptions 2.1–2.4 hold and* $0 < \eta \leq \eta_{\max}$. *Then,*

$$W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) \leq C_2^\star \eta^{1/4} d^{3/4}.$$

In particular, $C_2^\star$ comes from the contraction result of Eberle et al. (2019, Corollary 2.6) which might have exponential dependence in d as noted before.

Finally, the convergence of the last term follows from Theorem 3.1.

Theorem 4.3. *(Eberle et al., 2019, Gao et al., 2018) Let Assumptions 2.1, 2.2, 2.3, 2.4 hold. Then,*

$$W_2(\mathcal{L}(\zeta_t^\eta, Z_t^\eta), \overline{\pi}_\beta) \leq C_3^\star e^{-C_4^\star \eta t},$$

where $C_3^\star = \sqrt{\dot{C} \mathcal{W}_\rho(\mu_0, \nu_0)}$ *and* $C_4^\star = \dot{c}/2$.

Finally, considering Theorems 4.1, 4.2, and 4.3 together by putting $t = n$ leads to the full proof of our main result, namely, Theorem 2.1.

5 Proof of Theorem 2.2

The bound provided for the convergence to the target in W_2 distance can be used to obtain theoretical guarantees for the nonconvex optimization problem (12). In order to do so, we proceed by decomposing the error as follows

$$\mathbb{E}[U(\theta_n^\eta)] - U_\star = \underbrace{\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}[U(\theta_\infty)] - U_\star}_{\mathcal{T}_2},$$

where $\theta_\infty \sim \pi_\beta$. The following proposition presents a bound for $\mathcal{T}_1$ under our assumptions.

Proposition 5.1. *Under the assumptions of Theorem 2.1, we have,*

$$\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)] \leq \overline{C}_1^\star \eta^{1/2} d^{1/2} + \overline{C}_2^\star \eta^{1/4} d^{3/4} + \overline{C}_3^\star e^{-C_4^\star \eta n}, \tag{28}$$

where $\overline{C}_i^\star = C_i^\star (C_m \bar{L}_1 + h_0)$ *for* $i = 1, 2, 3$ *and* $C_m = \max(C_\theta^c, C_\theta)$.

Next, we bound the second term $\mathcal{T}_2$ as follows. This result is fairly standard in the literature (see, e.g., Raginsky et al. (2017), Gao et al. (2018), Chau and Rasonyi (2019)).

Proposition 5.2. *(Raginsky et al., 2017) Under the assumptions of Theorem 2.1, we have*

$$\mathbb{E}[U(\theta_\infty)] - U_\star \leq \frac{d}{2\beta} \log\left(\frac{e\overline{L}_1}{a}\left(\frac{b\beta}{d} + 1\right)\right).$$

Merging Props. 5.1 and 5.2 leads to the bound given in Theorem 2.2 which completes our proof.

6 Applications

In this section, we present two applications of our theory to machine learning problems. First, we show that the SGHMC can be used to sample from the posterior probability measure and can be used for scalable Bayesian inference. We also note that our assumptions hold in a practical setting of Bayesian logistic regression, as opposed to previous results. Secondly, we provide an improved generalization bound for empirical risk minimization using the SGHMC.

6.1 Convergence rates for scalable Bayesian inference

Consider a prior distribution $\pi_0(\theta)$ and a likelihood function $\mathsf{p}(y_i|\theta)$ for a sequence of data points $\{y_i\}_{i=1}^M$ where M is the dataset size. Often, one is interested in sampling from the posterior probability distribution

$$p(\theta|y_{1:M})\mathrm{d}\theta \propto \pi_0(\theta) \prod_{i=1}^{M} \mathsf{p}(y_i|\theta)\mathrm{d}\theta.$$

This is a sampling problem of the form (1). The SGHMC is an MCMC method to sample from the posterior measure π and, therefore, explicit convergence rates provides a guarantee for the sampling procedure. To see this, note that the underdamped Langevin SDE

$$\begin{aligned}\mathrm{d}V_t &= -\gamma V_t \mathrm{d}t + \nabla \log p(\theta_t|y_{1:M})\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}B_t,\\ \mathrm{d}\theta_t &= V_t \mathrm{d}t,\end{aligned}$$

converges to the extended target

$$\overline{\pi}(\mathrm{d}\theta, \mathrm{d}v) \propto \exp\left(-\frac{1}{2}\|v\|^2 + \log p(\theta|y_{1:M})\right)\mathrm{d}\theta\mathrm{d}v.$$

One can see that θ-marginal of $\overline{\pi}$ is precisely $p(\theta|y_{1:M})$, hence the underdamped Langevin SDE samples from the posterior. Therefore, the SGHMC can be used for sampling when the gradient of the target is only accessible with noise.

We note that, our setting specifically applies to cases where M is too large. More precisely, note that we have

$$\begin{aligned}h(\theta) &= -\nabla \log p(\theta|y_{1:M}),\\ &= -\nabla \log \pi_0(\theta) - \sum_{i=1}^{M} \nabla \log \mathsf{p}(y_i|\theta).\end{aligned} \tag{29}$$

When M is too large, evaluating $h(\theta)$ is impractical. However, one can estimate the sum in the last term of (29) in an unbiased way. To be precise, consider random indices $i_1, \ldots, i_K \sim \{1, ..., M\}$ uniformly, then one can construct a stochastic gradient by using $\mathbf{u} = \{y_{i_1}, \ldots, y_{i_K}\}$

$$H(\theta, \mathbf{u}) = -\log \pi_0(\theta) - \frac{M}{K} \sum_{k=1}^{K} \nabla \log \mathsf{p}(y_{i_k}|\theta).$$

Then, we have the simple corollary for Bayesian inference as a consequence of Theorem 2.1.

Corollary 6.1. *Assume that the log-posterior density* $\log p(\theta|y_{1:M})$*, its gradient, and stochastic gradient* $H(\theta, \cdot)$ *satisfy the Assumptions 2.1–2.4. Then,*

$$W_2(\mathcal{L}(\theta_n), p(\theta|y_{1:M})) \leq C_1^\star \eta^{1/2} d^{1/2} + C_2^\star \eta^{1/4} d^{3/4} + C_3^\star e^{-C_4^\star \eta n}.$$

where $C_1^\star, C_2^\star, C_3^\star, C_4^\star$ *are finite constants.*

This setting becomes practical under our assumptions, e.g., for the Bayesian logistic regression example. Consider the Gaussian mixture prior

$$\pi_0(\theta) \propto \exp(-f_0(\theta)) = e^{-|\theta - m|^2/2} + e^{-|\theta + m|^2/2},$$

where $m \in \mathbb{R}^d$ and the likelihood

$$\mathsf{p}(\mathbf{z}_i|\theta) = (1/(1 + e^{-z_i^\top \theta}))^{y_i}(1 - 1/(1 + e^{-z_i^\top \theta}))^{1-y_i},$$

for $\theta \in \mathbb{R}^d$ and $\mathbf{z}_i = (z_i, y_i)$. Then, it is shown by Zhang et al. (2019) that the stochastic gradient $H(\theta, \mathbf{u})$ for a mini-batch in this case satisfies assumptions 2.1–2.4. In particular, our theoretical guarantee in Theorem 2.1 and Corollary 6.1 apply to the Bayesian logistic regression case.

6.2 A generalization bound for machine learning

Leveraging standard results in machine learning literature, e.g., Raginsky et al. (2017), we can prove a generalization bound for the empirical risk minimization problem. Note that, many problems in machine learning can be written as a finite-sum minimization problem as

$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^d} U(\theta) := \frac{1}{M}\sum_{i=1}^{M} f(\theta, z_i). \tag{30}$$

Applying the result of Theorem 2.2, one can get a convergence guarantee on $\mathbb{E}[U(\theta_k^\eta)] - U_\star$. However, this does not account for the so-called *generalization error*. Note that, one can see the cost function in (30) as an empirical risk (expectation) minimization problem where the risk is given by

$$\mathbb{U}(\theta) := \int f(\theta, z) P(\mathrm{d}z) = \mathbb{E}[f(\theta, Z)],$$

where $Z \sim P(\mathrm{d}z)$ is an unknown probability measure where the real-world data is sampled from. Therefore, in order to bound the generalization error, one needs to bound the error $\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{U}_\star$.

The generalization error can be decomposed as

$$\begin{aligned}\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{U}_\star = & \underbrace{\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{E}[\mathbb{U}(\theta_\infty)]}_{\mathcal{B}_1} \\ & + \underbrace{\mathbb{E}[\mathbb{U}(\theta_\infty)] - \mathbb{E}[U(\theta_\infty)]}_{\mathcal{B}_2} \\ & + \underbrace{\mathbb{E}[U(\theta_\infty)] - \mathbb{U}_\star}_{\mathcal{B}_3}.\end{aligned}$$

In what follows, we present a series of results bounding the terms $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$. By using the results about Gibbs distributions presented in Raginsky et al. (2017), one can prove the following result, bounding $\mathcal{B}_1$.

Proposition 6.1. *Under the assumptions of Theorem 2.1, we obtain*

$$\mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{E}[\mathbb{U}(\theta_\infty)] \le \overline{C}_1^\star \eta^{1/2} d^{1/2} + \overline{C}_2^\star \eta^{1/4} d^{3/4} + \overline{C}_3^\star e^{-C_4^\star \eta n},$$

where $\overline{C}_i^\star = C_i^\star (C_m \bar{L}_1 + h_0)$ *for* $i = 1, 2, 3$ *and* $C_m = \max(C_\theta^c, C_\theta)$.

The proof of Proposition 6.1 is similar to the proof of Proposition 5.1 and indeed the rates of these results match.

Next, we seek a bound for the term $\mathcal{B}_2$. In order to be able to prove the following result about the stability of the Gibbs algorithm, we assume that Assumption 2.3 and Assumption 2.4 hold uniformly in x, as required by previous works, see, e.g., Raginsky et al. (2017), Gao et al. (2018), Chau and Rasonyi (2019).

Proposition 6.2. *(Raginsky et al., 2017) Assume that Assumptions 2.1, 2.2 hold and Assumptions 2.3 and 2.4 hold uniformly in x, i.e.,*

$$|H(\theta, x)| \le L_1' |\theta|^2 + B_1.$$

Then,

$$\mathbb{E}[\mathbb{U}(\theta_\infty)] - \mathbb{E}[U(\theta_\infty)] \le \frac{4\beta c_{\mathrm{LS}}}{M}\left(\frac{L_1'}{a}(b + d/\beta) + B_1\right),$$

where c_{LS} is the constant of the logarithmic Sobolev inequality.

Finally, let $\Theta^\star \in \arg\min_{\theta\in\mathbb{R}} \mathbb{U}(\theta)$. We note that $\mathcal{B}_3$ is bounded trivially as

$$\begin{aligned} \mathbb{E}[U(\theta_\infty)] - \mathbb{U}_\star &= \mathbb{E}[U(\theta_\infty) - U_\star] + \mathbb{E}[U_\star - U(\Theta^\star)] \\ &\leq \mathbb{E}[U(\theta_\infty) - U_\star], \\ &\leq \frac{d}{2\beta}\log\left(\frac{e\overline{L}_1}{a}\left(\frac{b\beta}{d}+1\right)\right), \end{aligned} \tag{31}$$

which follows from the proof of Proposition 5.2. Finally, Proposition 6.1, Proposition 6.2 and (31) leads to the following generalization bound presented as a corollary.

Corollary 6.2. *Under the setting of Proposition 6.2, we obtain the generalization bound for the SGHMC,*

$$\begin{aligned} \mathbb{E}[\mathbb{U}(\theta_n^\eta)] - \mathbb{U}_\star \leq \overline{C}_1^\star \eta^{1/2} d^{1/2} + \overline{C}_2^\star \eta^{1/4} d^{3/4} + \overline{C}_3^\star e^{-C_4^\star \eta n} + \frac{4\beta c_{\mathrm{LS}}}{M}\left(\frac{L_1'}{a}(b + d/\beta) + B_1\right) \\ + \frac{d}{2\beta}\log\left(\frac{e\overline{L}_1}{a}\left(\frac{b\beta}{d}+1\right)\right). \end{aligned} \tag{32}$$

We note that this generalization bound improves that of Raginsky et al. (2017), Gao et al. (2018), Chau and Rasonyi (2019) due to our improved W_2 bound which is reflected in Theorem 2.2 and, consequently, Proposition 6.1. In particular, while the generalization bounds of Raginsky et al. (2017) and Gao et al. (2018) grow with the number of iterations and require careful tuning between the step-size and the number of iterations, our bound decreases with the number of iterations n. We also note that our bound improves that of Chau and Rasonyi (2019), similar to the W_2 bound.

7 Conclusions

We have analyzed the convergence of the SGHMC recursions (6)–(7) to the extended target measure $\overline{\pi}_\beta$ in Wasserstein-2 distance which implies the convergence of the law of the iterates $\mathcal{L}(\theta_n^\eta)$ to the target measure π_β in W_2. We have proved that the error bound scales like $\mathcal{O}(\eta^{1/4})$ where η is the step-size. This improves the existing bounds for the SGHMC significantly which are either growing with the number of iterations or include constants cannot be made to vanish by decreasing the step-size η. This bound on sampling from π_β enables us to prove a stochastic global optimization result when $(\theta_n^\eta)_{n\in\mathbb{N}}$ is viewed as an output of a nonconvex optimizer. We have shown that our results provide convergence rates for scalable Bayesian inference and we have particularized our results to the Bayesian logistic regression. Moreover, we have shown that our improvement of W_2 bounds are reflected in improved generalization bounds for the SGHMC.

References

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *29th International Conference on Machine Learning, ICML 2012*, pages 1591–1598, 2012.

M. Barkhagen, N. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *arXiv preprint arXiv:1812.02709*, 2018.

K. Binder, D. Heermann, L. Roelofs, A. J. Mallinckrodt, and S. McKay. Monte Carlo simulation in statistical physics. *Computers in Physics*, 7(2):156–157, 1993.

N. Brosse, A. Durmus, and E. Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

H. N. Chau and M. Rasonyi. Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Learning. *arXiv preprint arXiv:1903.10328*, 2019.

H. N. Chau, C. Kumar, M. Rásonyi, and S. Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23:217–244, 2019a.

N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint arXiv:1905.13142*, 2019b.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference On Learning Theory*, pages 300–323, 2018b.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

A. Duncan, N. Nüsken, and G. Pavliotis. Using perturbed underdamped langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, 169(6):1098–1131, 2017.

A. Durmus, E. Moulines, et al. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Durmus, E. Moulines, et al. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization: Non-Asymptotic Performance Bounds and Momentum-Based Acceleration. *arXiv preprint arXiv:1809.04618*, 2018.

S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

J. M. Haile. *Molecular dynamics simulation: elementary methods*, volume 1. Wiley New York, 1992.

C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

Y.-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett, and M. I. Jordan. Is There an Analog of Nesterov Acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

M. B. Majka, A. Mijatović, and L. Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *arXiv preprint arXiv:1808.07105*, 2018.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

G. O. Roberts and O. Stramer. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

G. O. Roberts, R. L. Tweedie, et al. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

S. Sabanis and Y. Zhang. Higher order Langevin Monte Carlo algorithm. *Electronic Journal of Statistics*, 13(2):3805–3850, 2019.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

Y. Zhang, Ö. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *arXiv preprint arXiv:1910.02008*, 2019.

D. Zou, P. Xu, and Q. Gu. Stochastic gradient Hamiltonian monte carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems*, pages 3830–3841, 2019.

Supplementary Document

A Proofs of preliminary results

A.1 Additional lemmata

We first prove the following lemma adapted from Raginsky et al. (2017).

Lemma A.1. *For all $\theta \in \mathbb{R}^d$,*

$$\frac{a}{3}|\theta|^2 - \frac{b}{2}\log 3 \leq U(\theta) \leq u_0 + \frac{\bar{L}_1}{2}|\theta|^2 + \bar{L}_1 h_0 |\theta|.$$

where $u_0 = U(0)$ and $\bar{L}_1 = L_1 \mathbb{E}[(1+|X_0|)^\rho]$.

Proof. We start by writing that

$$\begin{aligned} U(\theta) - U(0) &= \int_0^1 \langle \theta, h(t\theta) \rangle \mathrm{d}t, \\ &\leq \int_0^1 |\theta||h(t\theta)|\mathrm{d}t, \\ &\leq \int_0^1 |\theta|\bar{L}_1(t|\theta| + h_0)\mathrm{d}t. \end{aligned}$$

from Remark (10) where $\bar{L}_1 = L_1\mathbb{E}[(1+|X_0|)^\rho]$. This in turn leads to

$$U(\theta) \leq u_0 + \frac{\bar{L}_1}{2}|\theta|^2 + \bar{L}_1 h_0|\theta|.$$

where $u_0 = U(0)$. Next, we prove the lower bound. To this end, take $c \in (0,1)$ and write

$$\begin{aligned} U(\theta) &= U(c\theta) + \int_c^1 \langle \theta, h(t\theta)\rangle \mathrm{d}t, \\ &\geq \int_c^1 \frac{1}{t}\langle t\theta, h(t\theta)\rangle \mathrm{d}t, \\ &\geq \int_c^1 \frac{1}{t}\left(a|t\theta|^2 - b\right)\mathrm{d}t, \\ &= \frac{a(1-c^2)}{2}|\theta|^2 + b\log c. \end{aligned}$$

Taking $c = 1/\sqrt{3}$ leads to the bound. □

Lemma A.2. *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be sigma-algebras. Consider two $\mathbb{R}^d$-valued random vectors, denoted X, Y, in L^p with $p \geq 1$ such that Y is measurable w.r.t. $\mathcal{H} \vee \mathcal{G}$. Then,*

$$\mathbb{E}^{1/p}[|X - \mathbb{E}[X|\mathcal{H}\vee\mathcal{G}]|^p|\mathcal{G}] \leq 2\mathbb{E}^{1/p}[|X-Y|^p|\mathcal{G}].$$

Proof. See Lemma 6.1 in Chau et al. (2019a). □

Lemma A.3. *Let Assumption 2.1, 2.2, 2.3 and 2.4 hold. For any $k = 1, \ldots, K+1$ where $K+1 \leq T$, we obtain*

$$\sup_{n\in\mathbb{N}}\sup_{t\in[nT,(n+1)T]} \mathbb{E}\left[\left|h(\bar{\zeta}_t^{\eta,n}) - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2\right] \leq \sigma_H$$

where

$$\sigma_H := 8L_2^2\sigma_Z(1+C_\zeta) < \infty,$$

where $\sigma_Z = \mathbb{E}[(1+|X_0|+|\mathbb{E}[X_0]|)^{2\rho}|X_0 - \mathbb{E}[X_0]|^2] < \infty$.

Proof. Let $\mathcal{H}_t = \mathcal{F}_\infty^\eta \vee \mathcal{G}_{\lfloor t \rfloor}$. Following Zhang et al. (2019), we obtain

$$\begin{aligned} &\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\eta,n}) - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\eta,n}) - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2 \middle| \mathcal{H}_{nT}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left|\mathbb{E}\left[H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\middle|\mathcal{H}_{nT}\right] - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2 \middle| \mathcal{H}_{nT}\right]\right] \\ &\leq 4\mathbb{E}\left[\mathbb{E}\left[\left|H(\bar{\zeta}_t^{\eta,n}, X_{nT+k}) - H(\bar{\zeta}_t^{\eta,n}, \mathbb{E}[X_{nT+k}|\mathcal{H}_{nT}])\right|^2 \middle| \mathcal{H}_{nT}\right]\right] \\ &\leq 4L_2^2\sigma_Z\mathbb{E}\left[(1+|\bar{\zeta}_t^{\eta,n}|)^2\right], \end{aligned}$$

where the first inequality holds due to Lemma A.2 and

$$\sigma_Z = \mathbb{E}[(1+|X_0|+|\mathbb{E}[X_0]|)^{2\rho}|X_0-\mathbb{E}[X_0]|^2].$$

Then, by using Lemma 3.4, we obtain

$$\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\eta,n}) - H(\bar{\zeta}_t^{\eta,n}, X_{nT+k})\right|^2\right] \leq 8L_2^2\sigma_Z + 8L_2^2\sigma_Z C_\zeta.$$

□

Lemma A.4. *Under Assumptions 2.1–2.4*

$$\mathbb{E}\left[\left|\overline{V}^\eta_{\lfloor t\rfloor} - \overline{V}^\eta_t\right|^2\right] \leq \sigma_V\eta,$$

where

$$\sigma_V = 4\eta\gamma^2 C_v + 4\eta(\widetilde{L}_1 C_\theta + \widetilde{C}_1) + 4\gamma\beta^{-1}d.$$

Moreover,

$$\mathbb{E}\left[\left|\overline{\theta}^\eta_{\lfloor t\rfloor} - \overline{\theta}^\eta_t\right|^2\right] \leq \eta^2 C_v.$$

Proof. Note that for any t, we have

$$\overline{V}^\eta_t = \overline{V}^\eta_{\lfloor t\rfloor} - \eta\gamma\int_{\lfloor t\rfloor}^t \overline{V}^\eta_{\lfloor s\rfloor}\mathrm{d}s - \eta\int_{\lfloor t\rfloor}^t H(\theta_{\lfloor s\rfloor}, X_{\lceil s\rceil})\mathrm{d}s + \sqrt{2\gamma\eta\beta^{-1}}(B^\eta_t - B^\eta_{\lfloor t\rfloor}).$$

We therefore obtain

$$\begin{aligned}\mathbb{E}\left[\left|\overline{V}^\eta_{\lfloor t\rfloor} - \overline{V}^\eta_t\right|^2\right] &= \mathbb{E}\left[\left|-\eta\gamma\int_{\lfloor t\rfloor}^t \overline{V}^\eta_{\lfloor s\rfloor}\mathrm{d}s - \eta\int_{\lfloor t\rfloor}^t H(\theta_{\lfloor s\rfloor}, X_{\lceil s\rceil})\mathrm{d}s + \sqrt{2\gamma\eta\beta^{-1}}(B^\eta_t - B^\eta_{\lfloor t\rfloor})\right|^2\right] \\ &\leq 4\eta^2\gamma^2 C_v + 4\eta^2(\widetilde{L}_1 C_\theta + \widetilde{C}_1) + 4\gamma\eta\beta^{-1}d.\end{aligned}$$

Next, we write

$$\overline{\theta}^\eta_t = \overline{\theta}^\eta_{\lfloor t\rfloor} + \int_{\lfloor t\rfloor}^t \eta\overline{V}^\eta_{\lfloor\tau\rfloor}\mathrm{d}\tau,$$

which implies

$$\mathbb{E}\left[\left|\overline{\theta}^\eta_t - \overline{\theta}^\eta_{\lfloor t\rfloor}\right|^2\right] \leq \eta^2 C_v.$$

□

A.2 Proofs of the preliminary results

A.2.1 Proof of Lemma 3.1

Choose $\lambda > 0$ such that

$$\lambda = \min\left\{\frac{1}{4}, \frac{a}{\bar{L}_1 + 2\bar{L}_1 b + \frac{\gamma^2}{2}}\right\}.$$

Using Assumption 2.4, we obtain

$$\begin{aligned}\langle\theta, h(\theta)\rangle &\geq a|\theta|^2 - b, \\ &= 2\lambda\left(\frac{\bar{L}_1}{2} + \bar{L}_1 b + \frac{\gamma^2}{4}\right)|\theta|^2 - b, \\ &\geq 2\lambda\left(U(\theta) - u_0 - \bar{L}_1 b|\theta| + \bar{L}_1 b|\theta|^2 + \frac{\gamma^2}{4}|\theta|^2\right) - b, \\ &\geq 2\lambda\left(U(\theta) - u_0 - \bar{L}_1 b + \frac{\gamma^2}{4}|\theta|^2\right) - b,\end{aligned}$$

where the third line follows from Lemma A.1 and the last line follows from the inequality $|x| \leq 1 + |x|^2$. Consequently, we obtain

$$\langle\theta, h(\theta)\rangle \geq 2\lambda\left(U(\theta) + \frac{\gamma^2}{4}|\theta|^2\right) - 2A_c/\beta$$

where

$$A_c = \frac{\beta}{2}(b + 2\lambda u_0 + 2\lambda\bar{L}_1 b),$$

which proves the claim.

A.2.2 Proof of Lemma 3.3

For this proof, we use the Lyapunov function defined by Eberle et al. (2019) and follow a similar proof presented in Gao et al. (2018). We first define the Lyapunov function as

$$\mathcal{V}(\theta, v) = \beta U(\theta) + \frac{\beta}{4}\gamma^2 \left(\|\theta + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|\theta\|^2 \right).$$

Next, we will use this Lyapunov function to show that the second moments of the processes $(V_n^\eta)_{n\in\mathbb{N}}$ and $(\theta_n^\eta)_{n\in\mathbb{N}}$ are finite.

We start by defining

$$M_2(k) = \mathbb{E}\mathcal{V}(\theta_k^\eta, V_k^\eta)/\beta = \mathbb{E}\left[U(\theta_k^\eta) + \frac{\gamma^2}{4}\left(|\theta_k^\eta + \gamma^{-1}V_k^\eta|^2 + |\gamma^{-1}V_k^\eta|^2 - \lambda|\theta_k^\eta|^2 \right) \right]. \tag{33}$$

Recall our discrete-time recursions (6)–(7)

$$\begin{aligned} V_{k+1}^\eta &= (1 - \eta\gamma)V_k^\eta - \eta H(\theta_k^\eta, X_{k+1}) + \sqrt{2\eta\gamma\beta^{-1}}\xi_{k+1}, \\ \theta_{k+1}^\eta &= \theta_k^\eta + \eta V_k^\eta, \qquad \theta_0^\eta = \theta_0, \qquad V_0^\eta = v_0, \end{aligned}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Normal random variables. Consequently, we have the equality

$$\begin{aligned} \mathbb{E}\left[|V_{k+1}^\eta|^2\right] &= \mathbb{E}\left[\left|(1-\gamma\eta)V_k^\eta - \eta H(\theta_k^\eta, X_{k+1})\right|^2\right] + 2\gamma\eta\beta^{-1}d, \\ &= (1-\gamma\eta)^2\mathbb{E}\left[\left|V_k^\eta\right|^2\right] - 2\eta(1-\gamma\eta)\mathbb{E}\left[\langle V_k^\eta, h(\theta_k^\eta)\rangle\right] + \eta^2\mathbb{E}\left[\left|H(\theta_k^\eta, X_{k+1})\right|^2\right] + 2\gamma\eta\beta^{-1}d, \end{aligned}$$

which immediately leads to

$$\mathbb{E}\left[|V_{k+1}^\eta|^2\right] \le (1-\gamma\eta)^2\mathbb{E}\left[\left|V_k^\eta\right|^2\right] - 2\eta(1-\gamma\eta)\mathbb{E}\left[\langle V_k^\eta, h(\theta_k^\eta)\rangle\right] + \eta^2\left[\widetilde{L}_1\mathbb{E}\left[\left|\theta_k^\eta\right|^2\right] + \widetilde{C}_1\right] + 2\gamma\eta\beta^{-1}d, \tag{34}$$

where

$$\widetilde{L}_1 = 2L_1^2 C_\rho \qquad \text{and} \qquad \widetilde{C}_1 = 4L_2^2 C_\rho + 4H_0^2. \tag{35}$$

Next, we note that

$$\mathbb{E}\left|\theta_{k+1}^\eta\right|^2 = \mathbb{E}|\theta_k^\eta|^2 + 2\eta\mathbb{E}\langle\theta_k^\eta, V_k^\eta\rangle + \eta^2\mathbb{E}\left|V_k^\eta\right|^2. \tag{36}$$

Recall $h := \nabla U$ and note also that

$$U(\theta_{k+1}^\eta) = U(\theta_k^\eta + \eta V_k^\eta) = U(\theta_k^\eta) + \int_0^1 \langle h(\theta_k^\eta + \tau\eta V_k^\eta), \eta V_k^\eta\rangle \mathrm{d}\tau,$$

which suggests

$$\begin{aligned} \left|U(\theta_{k+1}^\eta) - U(\theta_k^\eta) - \langle h(\theta_k^\eta), \eta V_k^\eta\rangle\right| &= \left|\int_0^1 \langle h(\theta_k^\eta + \tau\eta V_k^\eta), \eta V_k^\eta\rangle \mathrm{d}\tau\right|, \\ &\le \int_0^1 \left|h(\theta_k^\eta + \tau\eta V_k^\eta) - h(\theta_k^\eta)\right| \left|\eta V_k^\eta\right| \mathrm{d}\tau, \\ &\le \frac{1}{2}L_1 C_\rho \eta^2 \left|V_k^\eta\right|^2, \end{aligned}$$

where the second line follows from the Cauchy-Schwarz inequality and the final line follows from (9). Finally we obtain

$$\mathbb{E}U(\theta_{k+1}^\eta) - \mathbb{E}U(\theta_k^\eta) \le \eta\mathbb{E}\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{1}{2}L_1 C_\rho\eta^2\mathbb{E}\left|V_k^\eta\right|^2. \tag{37}$$

Next, we continue computing

$$\begin{aligned} \mathbb{E}\left|\theta_{k+1}^\eta + \gamma^{-1}V_{k+1}^\eta\right|^2 &= \mathbb{E}\left|\theta_k^\eta + \gamma^{-1}V_k^\eta - \eta\gamma^{-1}H(\theta_k^\eta, X_{k+1})\right|^2 + 2\gamma^{-1}\beta^{-1}\eta d, \\ &= \mathbb{E}\left|\theta_k^\eta + \gamma^{-1}V_k^\eta\right|^2 - 2\eta\gamma^{-1}\mathbb{E}\langle\theta_k^\eta + \gamma^{-1}V_k^\eta, h(\theta_k^\eta)\rangle \\ &+ \eta^2\gamma^{-2}\mathbb{E}\left|H(\theta_k^\eta, X_{k+1})\right|^2 + 2\gamma^{-1}\eta\beta^{-1}d, \\ &\le \mathbb{E}\left|\theta_k^\eta + \gamma^{-1}V_k^\eta\right|^2 - 2\eta\gamma^{-1}\mathbb{E}\langle\theta_k^\eta + \gamma^{-1}V_k^\eta, h(\theta_k^\eta)\rangle \\ &+ \eta^2\gamma^{-2}(\widetilde{L}_1\mathbb{E}\left|\theta_k^\eta\right|^2 + \widetilde{C}_1) + 2\gamma^{-1}\eta\beta^{-1}d. \end{aligned} \tag{38}$$

where $\widetilde{L}_1$ and $\widetilde{C}_1$ is defined as in (35). Next, combining (34), (36), (37), (38),

$$
\begin{aligned}
&M_2(k+1) - M_2(k) \\
&\quad = \mathbb{E}\left[U(\theta_{k+1}^\eta) - U(\theta_k^\eta)\right] + \frac{\gamma^2}{4}\left(\mathbb{E}\left|\theta_{k+1}^\eta + \gamma^{-1}V_{k+1}^\eta\right|^2 - \mathbb{E}\left|\theta_k^\eta + \gamma^{-1}V_k^\eta\right|^2\right) \\
&\quad + \frac{1}{4}\left(\mathbb{E}\left|V_{k+1}^\eta\right|^2 - \mathbb{E}\left|V_k^\eta\right|^2\right) - \frac{\gamma^2\lambda}{4}\left(\mathbb{E}\left|\theta_{k+1}^\eta\right|^2 - \mathbb{E}\left|\theta_k^\eta\right|^2\right), \\
&\quad \leq \eta\mathbb{E}\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{L_1 C_\rho \eta^2}{2}\mathbb{E}\left|V_k^\eta\right|^2 \\
&\quad + \frac{\gamma^2}{4}\left(-2\eta\gamma^{-1}\mathbb{E}\langle\theta_k^\eta + \gamma^{-1}V_k^\eta, h(\theta_k^\eta)\rangle + \eta^2\gamma^{-2}\left(\widetilde{L}_1\mathbb{E}\left|\theta_k^\eta\right|^2 + \widetilde{C}_1\right) + 2\gamma^{-1}\beta^{-1}\eta d\right) \\
&\quad + \frac{1}{4}\left((-2\gamma\eta + \gamma^2\eta^2)\mathbb{E}\left|V_k^\eta\right|^2 - 2\eta(1-\gamma\eta)\mathbb{E}\langle V_k^\eta, h(\theta_k^\eta)\rangle + \eta^2\left(\widetilde{L}_1\mathbb{E}|\theta_k^\eta|^2 + \widetilde{C}_1\right) + 2\gamma\eta\beta^{-1}d\right) \\
&\quad - \frac{\gamma^2\lambda}{4}\left(2\eta\mathbb{E}\langle\theta_k^\eta, V_k^\eta\rangle + \eta^2\mathbb{E}|V_k^\eta|^2\right), \\
&\quad = -\frac{\eta\gamma}{2}\mathbb{E}\langle\theta_k^\eta, h(\theta_k^\eta)\rangle + \frac{\gamma\eta^2}{2}\mathbb{E}\langle h(\theta_k^\eta), V_k^\eta\rangle + \left(\frac{L_1 C_\rho\eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4}\right)\mathbb{E}\left|V_k^\eta\right|^2 \\
&\quad + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\left|\theta_k^\eta\right|^2 - \frac{\gamma^2\eta\lambda}{2}\mathbb{E}\langle\theta_k^\eta, V_k^\eta\rangle + \frac{\widetilde{C}_1\eta^2}{2} + \gamma\eta\beta^{-1}d, \\
&\quad \leq -\eta\gamma\lambda\mathbb{E}U(\theta_k^\eta) - \frac{\lambda\gamma^3\eta}{4}\mathbb{E}|\theta_k^\eta|^2 + A_c\eta\gamma\beta^{-1} + \frac{\gamma\eta^2}{2}\mathbb{E}\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\left|\theta_k^\eta\right|^2 \\
&\quad + \left(\frac{L_1 C_\rho\eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4}\right)\mathbb{E}\left|V_k^\eta\right|^2 - \frac{\gamma^2\eta\lambda}{2}\mathbb{E}\langle\theta_k^\eta, V_k^\eta\rangle + \frac{\widetilde{C}_1\eta^2}{2} + \gamma\eta\beta^{-1}d.
\end{aligned}
$$

where the last line is obtained using (20). Next, using the fact that $0 < \lambda \leq 1/4$ and the form of the Lyapunov function (33), we obtain

$$
-\frac{\gamma}{2}\mathbb{E}\langle\theta_k^\eta, V_k^\eta\rangle \leq -M_2(k) + \mathbb{E}U(\theta_k^\eta) + \frac{\gamma^2}{4}\mathbb{E}|\theta_k^\eta|^2 + \frac{1}{2}\mathbb{E}|V_k^\eta|^2.
$$

Using this, we can obtain

$$
\begin{aligned}
&M_2(k+1) - M_2(k) \leq A_c\eta\gamma\beta^{-1} + \frac{\gamma\eta^2}{2}\mathbb{E}\langle h(\theta_k^\eta), V_k^\eta\rangle + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\left|\theta_k^\eta\right|^2 + \gamma\eta\beta^{-1}d \\
&+ \left(\frac{L_1 C_\rho\eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4} + \frac{\gamma\lambda\eta}{2}\right)\mathbb{E}\left|V_k^\eta\right|^2 - \frac{\gamma^2\eta\lambda}{2}\mathbb{E}\langle\theta_k^\eta, V_k^\eta\rangle + \frac{\widetilde{C}_1\eta^2}{2} - \gamma\lambda\eta M_2(k).
\end{aligned}
$$

Next, reorganizing and using $\langle a, b\rangle \leq (|a|^2 + |b|^2)/2$

$$
\begin{aligned}
M_2(k+1) &\leq (1 - \gamma\lambda\eta)M_2(k) + A_c\eta\gamma\beta^{-1} + \frac{\eta^2\widetilde{L}_1}{2}\mathbb{E}\left|\theta_k^\eta\right|^2 + \frac{\widetilde{C}_1\eta^2}{2} + \gamma\eta\beta^{-1}d \\
&+ \left(\frac{L_1 C_\rho\eta^2}{2} + \frac{\eta^2\gamma^2}{4} - \frac{\gamma\eta}{2} - \frac{\gamma^2\eta^2\lambda}{4} + \frac{\gamma\lambda\eta}{4} + \frac{\gamma\eta^2}{4}\right)\mathbb{E}\left|V_k^\eta\right|^2 + \frac{\gamma\eta^2}{4}\mathbb{E}|h(\theta_k^\eta)|^2, \\
&\leq (1 - \gamma\lambda\eta)M_2(k) + A_c\eta\gamma\beta^{-1} + \eta^2\left(\frac{\widetilde{L}_1}{2} + \frac{\gamma}{2}L_1 C_\rho^2\right)\mathbb{E}\left|\theta_k^\eta\right|^2 + \frac{\widetilde{C}_1\eta^2}{2} + \gamma\eta\beta^{-1}d \\
&+ \eta^2\left(\frac{L_1 C_\rho}{2} + \frac{\gamma^2}{4} - \frac{\gamma^2\lambda}{4} + \frac{\gamma}{4}\right)\mathbb{E}\left|V_k^\eta\right|^2 + \frac{\gamma\eta^2 h_0^2}{2},
\end{aligned}
$$

where the last inequality follows since $\lambda \leq 1/4$ and (10). We note that

$$
\mathcal{V}(\theta, v) \geq \max\left\{\frac{1}{8}(1-2\lambda)\beta\gamma^2|\theta|^2, \frac{\beta}{4}(1-2\lambda)|v|^2\right\},
$$

which implies by the definition of $M_2(k)$ that

$$
\begin{aligned}
M_2(k) &\geq \max\left\{\frac{1}{8}(1-2\lambda)\gamma^2\mathbb{E}|\theta_k^\eta|^2, \frac{1}{4}(1-2\lambda)\mathbb{E}|V_k^\eta|^2\right\}, \\
&\geq \frac{1}{16}(1-2\lambda)\gamma^2\mathbb{E}|\theta_k^\eta|^2 + \frac{1}{8}(1-2\lambda)\mathbb{E}|V_k^\eta|^2,
\end{aligned}
\tag{39}
$$

since $\max\{x, y\} \geq (x+y)/2$ for any $x, y > 0$. Therefore, we obtain

$$
M_2(k+1) \leq (1 - \gamma\lambda\eta + K_1\eta^2)M_2(k) + K_2\eta^2 + K_3\eta
$$

where

$$K_1 := \max\left\{\frac{\frac{L_1 C_\rho}{2}+\frac{\gamma^2}{4}-\frac{\gamma^2\lambda}{4}+\frac{\gamma}{4}}{\frac{1}{8}(1-2\lambda)}, \frac{\frac{\widetilde{L}_1}{2}+\frac{\gamma}{2}L_1 C_\rho^2}{\frac{1}{16}(1-2\lambda)\gamma^2}\right\}$$

and

$$K_2 = \frac{\widetilde{C_1}+\gamma h_0^2}{2} \quad \text{and} \quad K_3 = (A_c + d)\gamma\beta^{-1}.$$

For $0 < \eta \leq \min\left\{\frac{K_3}{K_2}, \frac{\gamma\lambda}{K_1}, \frac{1}{\gamma\lambda}\right\}$, we obtain

$$M_2(k+1) \leq (1-\gamma\lambda\eta)\, M_2(k) + 2K_3\eta$$

which implies

$$M_2(k) \leq M_2(0) + \frac{2}{\gamma\lambda}K_3.$$

Combining this with (39) gives the result.

A.3 Proof of Lemma 3.4

We recall that $\overline{\zeta}_t^{\eta,n}$ is the Langevin diffusion started at θ_{nT}^η and run until $t \in (nT, (n+1)T)$. First notice that Lemma 3.2 implies

$$\sup_{t\in(nT,(n+1)T)} \mathbb{E}|\zeta_t^{\eta,n}|^2 \leq \frac{\mathbb{E}\mathcal{V}(\theta_{nT}^\eta, V_{nT}^\eta) + \frac{2(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}, \tag{40}$$

which, noting that, $\mathbb{E}\mathcal{V}(\theta_{nT}^\eta, V_{nT}^\eta) = \beta M_2(nT)$, implies

$$\sup_{t\in(nT,(n+1)T)} \mathbb{E}|\zeta_t^{\eta,n}|^2 \leq \frac{\beta M_2(0) + \frac{4(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}.$$

Substituting $M_2(0)$ gives

$$\sup_{t\in(nT,(n+1)T)} \mathbb{E}|\zeta_t^{\eta,n}|^2 \leq C_\zeta := \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(\theta,v)\mu_0(\mathrm{d}\theta,\mathrm{d}v) + \frac{4(d+A_c)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}.$$

B Proofs of main results

B.1 Proof of Theorem 4.1

We note that

$$W_2(\mathcal{L}(\overline{\theta}_t^\eta, \overline{V}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})) = W_2(\mathcal{L}(\overline{\theta}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n})) + W_2(\mathcal{L}(\overline{V}_t^\eta), \mathcal{L}(\overline{Z}_t^{\eta,n})). \tag{41}$$

We first bound the first term of (41). We start by employing the synchronous coupling and obtain

$$\left|\overline{\theta}_t^\eta - \overline{\zeta}_t^{\eta,n}\right| \leq \eta\int_{nT}^t \left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right| \mathrm{d}s,$$

which implies

$$\begin{aligned}
\sup_{nT\leq u\leq t} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] &\leq \eta \sup_{nT\leq u\leq t}\int_{nT}^u \mathbb{E}\left[\left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right] \mathrm{d}s,\\
&= \eta\int_{nT}^t \mathbb{E}\left[\left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right] \mathrm{d}s,
\end{aligned}$$

Next, we write for any $t \in [nT, (n+1)T)$

$$\begin{aligned}
\left|\overline{V}_{\lfloor t\rfloor}^\eta - \overline{Z}_t^{\eta,n}\right| &\leq \left|\overline{V}_{\lfloor t\rfloor}^\eta - \overline{V}_t^\eta\right| + \left|\overline{V}_t^\eta - \overline{Z}_t^{\eta,n}\right|,\\
&\leq \left|\overline{V}_{\lfloor t\rfloor}^\eta - \overline{V}_t^\eta\right| + \left|-\gamma\eta\int_{nT}^t [\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}]\mathrm{d}s - \eta\int_{nT}^t [H(\overline{\theta}_{\lfloor s\rfloor}^\eta, X_{\lceil s\rceil}) - h(\overline{\zeta}_s^{\eta,n})]\mathrm{d}s\right|,\\
&\leq \left|\overline{V}_{\lfloor t\rfloor}^\eta - \overline{V}_t^\eta\right| + \gamma\eta\int_{nT}^t \left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|\mathrm{d}s + \eta\left|\int_{nT}^t [H(\overline{\theta}_{\lfloor s\rfloor}^\eta, X_{\lceil s\rceil}) - h(\overline{\zeta}_s^{\eta,n})]\mathrm{d}s\right|,\\
&\leq \left|\overline{V}_{\lfloor t\rfloor}^\eta - \overline{V}_t^\eta\right| + \gamma\eta\int_{nT}^t \left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|\mathrm{d}s\\
&+ \eta\int_{nT}^t \left|H(\overline{\theta}_{\lfloor s\rfloor}^\eta, X_{\lceil s\rceil}) - H(\overline{\zeta}_s^{\eta,n}, X_{\lceil s\rceil})\right|\mathrm{d}s + \eta\left|\int_{nT}^t [H(\overline{\zeta}_s^{\eta,n}, X_{\lceil s\rceil}) - h(\overline{\zeta}_s^{\eta,n})]\mathrm{d}s\right|,\\
&\leq \left|\overline{V}_{\lfloor t\rfloor}^\eta - \overline{V}_t^\eta\right| + \gamma\eta\int_{nT}^t \left|\overline{V}_{\lfloor s\rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|\mathrm{d}s\\
&+ \eta\int_{nT}^t L_1(1+X_{\lceil s\rceil})^\rho \left|\overline{\theta}_{\lfloor s\rfloor}^\eta - \overline{\zeta}_s^{\eta,n}\right|\mathrm{d}s + \eta\left|\int_{nT}^t [H(\overline{\zeta}_s^{\eta,n}, X_{\lceil s\rceil}) - h(\overline{\zeta}_s^{\eta,n})]\mathrm{d}s\right|.
\end{aligned}$$

We take the squares of both sides and use $(a+b)^2 \leq a^2 + b^2$ twice to obtain

$$\begin{aligned}
\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right|^2 &\leq 4\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right|^2 + 4\gamma^2\eta^2 \left(\int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right| \mathrm{d}s\right)^2 \\
&\quad + 4\eta^2 \left(\int_{nT}^t L_1(1+X_{\lceil s \rceil})^{\rho} \left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right| \mathrm{d}s\right)^2 + 4\eta^2 \left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2 \\
&\leq 4\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right|^2 + 4\gamma^2\eta \int_{nT}^t \left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2 \mathrm{d}s \\
&\quad + 4\eta L_1^2 \int_{nT}^t (1+X_{\lceil s \rceil})^{2\rho} \left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2 \mathrm{d}s + 4\eta^2 \left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2.
\end{aligned}$$

Taking expectations of both sides, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right|^2\right] &\leq 4\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{V}^{\eta}_t\right|^2\right] + 4\gamma^2\eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2\right] \mathrm{d}s \\
&\quad + 4\eta L_1^2 C_{\rho} \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2\right] \mathrm{d}s + 4\eta^2 \mathbb{E}\left[\left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2\right], \\
&\leq 4\sigma_V\eta + 4\gamma^2\eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2\right] \mathrm{d}s \\
&\quad + 4\eta L_1^2 C_{\rho} \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2\right] \mathrm{d}s + 4\eta^2 \mathbb{E}\left[\left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2\right].
\end{aligned}$$

By applying Grönwall's lemma, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor t \rfloor} - \overline{Z}^{\eta,n}_t\right|^2\right] &\leq 4c_1\sigma_V\eta + 4\eta c_1 L_1^2 C_{\rho} \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s \rfloor} - \overline{\zeta}^{\eta,n}_s\right|^2\right] \mathrm{d}s \\
&\quad + 4c_1\eta^2 \mathbb{E}\left[\left|\int_{nT}^t [H(\overline{\zeta}^{\eta,n}_s, X_{\lceil s \rceil}) - h(\overline{\zeta}^{\eta,n}_s)]\mathrm{d}s\right|^2\right],
\end{aligned}$$

where $c_1 = \exp(4\gamma^2)$ since $\eta T \leq 1$. Next, we write

$$\begin{aligned}
\sup_{nT \leq u \leq t} \mathbb{E}\left[\left|\overline{\theta}^{\eta}_u - \overline{\zeta}^{\eta,n}_u\right|^2\right] &\leq \eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}^{\eta}_{\lfloor s \rfloor} - \overline{Z}^{\eta,n}_s\right|^2\right] \mathrm{d}s, \\
&\leq 4c_1\eta\sigma_V + 4\eta^2 L_1^2 C_{\rho} \int_{nT}^t \int_{nT}^s \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s' \rfloor} - \overline{\zeta}^{\eta,n}_{s'}\right|^2\right] \mathrm{d}s'\mathrm{d}s \\
&\quad + 4c_1\eta^3 \int_{nT}^t \mathbb{E}\left[\left|\int_{nT}^s [H(\overline{\zeta}^{\eta,n}_{s'}, X_{\lceil s' \rceil}) - h(\overline{\zeta}^{\eta,n}_{s'})]\mathrm{d}s'\right|^2\right] \mathrm{d}s, \\
&\leq 4c_1\eta\sigma_V + 4\eta L_1^2 C_{\rho} \sup_{nT \leq s \leq t} \int_{nT}^s \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s' \rfloor} - \overline{\zeta}^{\eta,n}_{s'}\right|^2\right] \mathrm{d}s' \\
&\quad + 4c_1\eta^3 \int_{nT}^t \mathbb{E}\left[\left|\int_{nT}^s [H(\overline{\zeta}^{\eta,n}_{s'}, X_{\lceil s' \rceil}) - h(\overline{\zeta}^{\eta,n}_{s'})]\mathrm{d}s'\right|^2\right] \mathrm{d}s. \qquad (42)
\end{aligned}$$

First, we bound the supremum term (i.e. the second term of (42)) as

$$\begin{aligned}
\sup_{nT \leq s \leq t} \int_{nT}^s \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s' \rfloor} - \overline{\zeta}^{\eta,n}_{s'}\right|^2\right] \mathrm{d}s' &= \int_{nT}^t \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor s' \rfloor} - \overline{\zeta}^{\eta,n}_{s'}\right|^2\right] \mathrm{d}s' \\
&\leq \int_{nT}^t \sup_{nT \leq u \leq s'} \mathbb{E}\left[\left|\overline{\theta}^{\eta}_{\lfloor u \rfloor} - \overline{\zeta}^{\eta,n}_u\right|^2\right] \mathrm{d}s' \\
&\leq \int_{nT}^t \sup_{nT \leq u \leq s'} \mathbb{E}\left[\left|\overline{\theta}^{\eta}_u - \overline{\zeta}^{\eta,n}_u\right|^2\right] \mathrm{d}s'. \qquad (43)
\end{aligned}$$

Next, we bound the last term of (42) by partitioning the integral. Assume that $nT + K \leq s \leq t \leq nT + K + 1$ where $K + 1 \leq T$. Thus we can write

$$\left|\int_{nT}^s \left[h(\bar{\zeta}^{\eta,n}_{s'}) - H(\bar{\zeta}^{\eta,n}_{s'}, X_{\lceil s' \rceil})\right] \mathrm{d}s'\right| = \left|\sum_{k=1}^K I_k + R_K\right|$$

where

$$I_k = \int_{nT+(k-1)}^{nT+k} [h(\bar{\zeta}^{\eta,n}_{s'}) - H(\bar{\zeta}^{\eta,n}_{s'}, X_{nT+k})]\mathrm{d}s' \qquad \text{and} \qquad R_K = \int_{nT+K}^s [h(\bar{\zeta}^{\eta,n}_{s'}) - H(\bar{\zeta}^{\eta,n}_{s'}, X_{nT+K+1})]\mathrm{d}s'.$$

Taking squares of both sides

$$\left|\sum_{k=1}^K I_k + R_K\right|^2 = \sum_{k=1}^K |I_k|^2 + 2\sum_{k=2}^K \sum_{j=1}^{k-1} \langle I_k, I_j \rangle + 2\sum_{k=1}^K \langle I_k, R_K \rangle + |R_K|^2,$$

Finally, it remains to take the expectations of both sides. We begin by defining the filtration $\mathcal{H}_s = \mathcal{F}_\infty^\eta \vee \mathcal{G}_{\lfloor s \rfloor}$ and note that for any $k = 2, \dots, K, j = 1, \dots, k-1$,

$$
\begin{aligned}
&\mathbb{E}\langle I_k, I_j \rangle \\
&= \mathbb{E}\left[\mathbb{E}[\langle I_k, I_j\rangle | \mathcal{H}_{nT+j}]\right], \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \int_{nT+(k-1)}^{nT+k} [H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k}) - h(\bar{\zeta}_{s'}^{\eta,n})]\mathrm{d}s', \int_{nT+(j-1)}^{nT+j} [H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+j}) - h(\bar{\zeta}_{s'}^{\eta,n})]\mathrm{d}s' \right\rangle \middle| \mathcal{H}_{nT+j}\right]\right], \\
&= \mathbb{E}\left[\left\langle \int_{nT+(k-1)}^{nT+k} \mathbb{E}\left[H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k}) - h(\bar{\zeta}_{s'}^{\eta,n}) \middle| \mathcal{H}_{nT+j}\right] \mathrm{d}s', \int_{nT+(j-1)}^{nT+j} [H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+j}) - h(\bar{\zeta}_{s'}^{\eta,n})]\mathrm{d}s' \right\rangle\right], \\
&= 0.
\end{aligned}
$$

By the same argument $\mathbb{E}\langle I_k, R_K \rangle = 0$ for all $1 \le k \le K$. Therefore,

$$
\begin{aligned}
\int_{nT}^t \mathbb{E}&\left[\left|\int_{nT}^s [H(\overline{\zeta}_{s'}^{\eta,n}, X_{\lceil s' \rceil}) - h(\overline{\zeta}_{s'}^{\eta,n})]\mathrm{d}s'\right|^2\right] \mathrm{d}s \\
&= \int_{nT}^t \left[\sum_{k=1}^K \mathbb{E}\left[\left|\int_{nT+(k-1)}^{nT+k} [h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k})]\mathrm{d}s'\right|^2\right]\right] \mathrm{d}s \\
&+ \int_{nT}^t \mathbb{E}\left[\left|\int_{nT+K}^s [h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+K+1})]\mathrm{d}s'\right|^2\right] \mathrm{d}s \\
&\le \int_{nT}^t \left[\sum_{k=1}^K \int_{nT+(k-1)}^{nT+k} \mathbb{E}\left[\left|h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+k})\right|^2\right] \mathrm{d}s'\right] \mathrm{d}s \\
&+ \int_{nT}^t \int_{nT+K}^s \mathbb{E}\left[\left|h(\bar{\zeta}_{s'}^{\eta,n}) - H(\bar{\zeta}_{s'}^{\eta,n}, X_{nT+K+1})\right|^2\right] \mathrm{d}s'\mathrm{d}s \\
&\le T^2\sigma_H + T\sigma_H.
\end{aligned} \tag{44}
$$

Using (42), (43), and (44), we eventually obtain

$$
\begin{aligned}
\sup_{nT \le u \le t} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] &\le 4c_1\eta\sigma_V + 4\eta L_1^2 C_\rho \int_{nT}^t \sup_{nT \le u \le s'} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] \mathrm{d}s', \\
&+ 4c_1\eta^3(T^2\sigma_H + T\sigma_H), \\
&\le 4c_1\eta\sigma_V + 4\eta L_1^2 C_\rho \int_{nT}^t \sup_{nT \le u \le s'} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] \mathrm{d}s' \\
&+ 4c_1\eta\sigma_H + 4c_1\eta^2\sigma_H,
\end{aligned} \tag{45}
$$

since $\eta T \le 1$. Finally, applying Grönwall's inequality and using again $\eta T \le 1$ provides

$$
\sup_{nT \le u \le t} \mathbb{E}\left[\left|\overline{\theta}_u^\eta - \overline{\zeta}_u^{\eta,n}\right|^2\right] \le \exp(4L_1^2 C_\rho)(4c_1\sigma_V + 4c_1\sigma_H + 4c_1\eta\sigma_H)\eta,
$$

which implies that.

$$
W_2(\mathcal{L}(\overline{\theta}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n})) \le C_{1,1}^\star \sqrt{\eta} \tag{46}
$$

with $C_{1,1}^\star = \sqrt{\exp(4L_1^2 C_\rho)(4c_1\sigma_V + 4c_1\sigma_H + 4c_1\eta\sigma_H)}$. Note that $\sigma_V = \mathcal{O}(d)$ and $\sigma_H = \mathcal{O}(d)$ hence $C_{1,1}^\star = \mathcal{O}(\sqrt{d})$.

Next, we upper bound the second term of (41). To prove it, we write

$$
\left|\overline{V}_t^\eta - \overline{Z}_t^{\eta,n}\right| \le \left|\gamma\eta \int_{nT}^t \left[\overline{V}_{\lfloor s \rfloor}^\eta - \overline{Z}_s^{\eta,n}\right] \mathrm{d}s\right| + \eta \left|\int_{nT}^t \left[H(\overline{\theta}_{\lfloor s \rfloor}^\eta) - h(\overline{\zeta}_s^{\eta,n})\right] \mathrm{d}s\right|,
$$

which leads to

$$
\mathbb{E}\left[\left|\overline{V}_t^\eta - \overline{Z}_t^{\eta,n}\right|^2\right] \le 2\gamma^2\eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}_{\lfloor s \rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right] + 2\eta^2 \mathbb{E}\left[\left|\int_{nT}^t \left[H(\overline{\theta}_{\lfloor s \rfloor}^\eta) - h(\overline{\zeta}_s^{\eta,n})\right] \mathrm{d}s\right|^2\right].
$$

By similar arguments we have used for bounding the the first term, we obtain

$$
\mathbb{E}\left[\left|\overline{V}_t^\eta - \overline{Z}_t^{\eta,n}\right|^2\right] \le 2\gamma^2\eta \int_{nT}^t \mathbb{E}\left[\left|\overline{V}_{\lfloor s \rfloor}^\eta - \overline{Z}_s^{\eta,n}\right|^2\right] + \sigma_H\eta + \sigma_H\eta^2.
$$

Using the fact that the rhs is an increasing function of t and we obtain

$$
\sup_{nT \le u \le t} \mathbb{E}\left[\left|\overline{V}_u^\eta - \overline{Z}_u^{\eta,n}\right|^2\right] \le 2\gamma^2\eta \int_{nT}^t \sup_{nT \le u \le s} \mathbb{E}\left[\left|\overline{V}_u^\eta - \overline{Z}_u^{\eta,n}\right|^2\right] \mathrm{d}s + \sigma_H\eta + \sigma_H\eta^2,
$$

Applying Gronwall's lemma and $\eta T \le 1$ yields

$$
\sup_{nT \le u \le t} \mathbb{E}\left[\left|\overline{V}_u^\eta - \overline{Z}_u^{\eta,n}\right|^2\right] \le \exp(2\gamma^2)(\sigma_H\eta + \sigma_H\eta^2),
$$

which leads to

$$W_2(\mathcal{L}(\overline{V}_t^\eta, \overline{Z}_t^{\eta,n}) \leq C_{1,2}^\star \sqrt{\eta}, \tag{47}$$

where $C_{1,2}^\star = \sqrt{\exp(2\gamma^2)\sigma_H(1+\eta)}$. Note again that $\sigma_H = \mathcal{O}(d)$, hence $C_{1,2}^\star = \mathcal{O}(d^{1/2})$.

Therefore, combining (41), (46), (47), we obtain

$$W_2(\mathcal{L}(\overline{\theta}_t^\eta, \overline{V}_t^\eta), \mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n})) \leq C_1^\star d^{1/2}\eta^{1/2}.$$

B.2 Proof of Theorem 4.2

Triangle inequality implies that

$$\begin{aligned}
W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,n}, \overline{Z}_t^{\eta,n}), \mathcal{L}(\zeta_t^\eta, Z_t^\eta)) &\leq \sum_{k=1}^n W_2(\mathcal{L}(\overline{\zeta}_t^{\eta,k}, \overline{Z}_t^{\eta,k}), \mathcal{L}(\overline{\zeta}_t^{\eta,k-1}), \overline{Z}_t^{\eta,k-1})),\\
&= W_2(\mathcal{L}(\zeta_t^{kT,\theta_{kT}^\eta,V_{kT}^\eta,\eta}, Z_t^{kT,\theta_{kT}^\eta,V_{kT}^\eta,\eta}), \mathcal{L}(\zeta_t^{(k-1)T,\theta_{(k-1)T}^\eta,V_{(k-1)T}^\eta,\eta}), Z_t^{(k-1)T,\theta_{(k-1)T}^\eta,V_{(k-1)T}^\eta,\eta})),\\
&= W_2(\mathcal{L}(\zeta_t^{kT,\theta_{kT}^\eta,V_{kT}^\eta,\eta}, Z_t^{kT,\theta_{kT}^\eta,V_{kT}^\eta,\eta}), \mathcal{L}(\zeta_t^{kT,\zeta_{(k-1)T}^\eta,Z_{(k-1)T}^\eta,\eta}, Z_t^{kT,\zeta_{(k-1)T}^\eta,Z_{(k-1)T}^\eta,\eta})),\\
&\leq \sqrt{\dot{C}} \sum_{k=1}^n e^{-\eta\dot{c}(t-kT)/2} \sqrt{\mathcal{W}_\rho(\mathcal{L}(\theta_{kT}^\eta, V_{kT}^\eta), \mathcal{L}(\zeta_{kT}^{(k-1)T,\theta_{(k-1)T}^\eta,V_{(k-1)T}^\eta,\eta}, Z_{kT}^{(k-1)T,\theta_{(k-1)T}^\eta,V_{(k-1)T}^\eta,\eta})},\\
&\leq 3\max\{1+\alpha, \gamma^{-1}\}\sqrt{\dot{C}} \sum_{k=1}^n e^{-\eta\dot{c}(t-kT)/2} \sqrt{1+\varepsilon_c \mathbb{E}^{1/2}[\mathcal{V}^2(\theta_{kT}^\eta, V_{kT}^\eta)] + \varepsilon_c \mathbb{E}^{1/2}[\mathcal{V}^2(\overline{\zeta}_{kT}^{\eta,(k-1)T}, \overline{Z}_{kT}^{\eta,(k-1)T})]}\\
&\times \sqrt{W_2(\mathcal{L}(\theta_{kT}^\eta, V_{kT}^\eta), \mathcal{L}(\overline{\zeta}_{kT}^{\eta,(k-1)T}, \overline{Z}_{kT}^{\eta,(k-1)T}))},\\
&\leq C_2^\star \eta^{1/4} d^{3/4},
\end{aligned}$$

where the last two lines follow from Theorem 4.2 in Eberle et al. (2019) and Theorem 2.8, Lemma 5.4 and Lemma 5.5 in Chau and Rasonyi (2019), respectively.

B.3 Proof of Proposition 5.1

We denote $\pi_{n,\beta}^\eta := \mathcal{L}(\theta_n^\eta, V_n^\eta)$ and write

$$\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)] = \int_{\mathbb{R}^{2d}} U(\theta)\pi_{n,\beta}^\eta(\mathrm{d}\theta, \mathrm{d}v) - \int_{\mathbb{R}^{2d}} U(\theta)\pi_\beta(\mathrm{d}\theta, \mathrm{d}v).$$

Recall from (10), we have

$$|h(\theta)| \leq \overline{L}_1|\theta| + h_0.$$

Using the arguments in Raginsky et al. (2017) and Gao et al. (2018), we arrive at

$$\left|\int_{\mathbb{R}^{2d}} U(\theta)\pi_{n,\beta}^\eta(\mathrm{d}\theta, \mathrm{d}v) - \int_{\mathbb{R}^{2d}} U(\theta)\pi_\beta(\mathrm{d}\theta, \mathrm{d}v)\right| \leq (\overline{L}_1 C_m + h_0) W_2(\pi_{n,\beta}^\eta, \pi_\beta),$$

where,

$$C_m := \max\left(\int_{\mathbb{R}^{2d}} \|\theta\|^2 \pi_{n,\beta}^\eta(\mathrm{d}\theta, \mathrm{d}v), \int_{\mathbb{R}^{2d}} \|\theta\|^2 \pi_\beta(\mathrm{d}\theta, \mathrm{d}v)\right) = \max(C_\theta^c, C_\theta).$$

We therefore obtain using Theorem 2.1 that

$$\mathbb{E}[U(\theta_n^\eta)] - \mathbb{E}[U(\theta_\infty)] \leq (\overline{L}_1 C_m + h_0)\left(C_1^\star d^{1/2}\eta^{1/2} + C_2^\star \eta^{1/4} d^{1/2} + C_3^\star e^{-C_4^\star \eta n}\right).$$